

Towards Contextual and Structural Relevance Feedback in XML Retrieval

Lobna Hlaoua and Mohand Boughanem

IRIT-SIG, 118 route de Narbonne 31068 Toulouse Cedex 4, France

{hlaoua,bougha}@irit.fr

Abstract

XML Retrieval is a process whose objective is to give the most exhaustive and specific information for given query. Relevance Feedback in XML retrieval has been recently investigated it consists in considering both content and structural information extracted from elements judged relevant, in query reformulation. In this paper, we describe a preliminary approach to select the most expressive keywords and the most appropriate generative structure to be added to the user query.

1. Introduction

The Relevance Feedback (RF) is an interactive and evaluative process. It usually consists in enriching an initial request by adding terms extracted from documents judged as relevant by the user.

Recently, the new standards of document representation have appeared, in particular, XML (eXtensible Markup Language) developed by W3C [10]. By exploring the characteristics of these new standard, traditional Information Retrieval (IR) that treats a document like only one atomic unit, has been extended to better manage this kind of documents. Indeed due to the structure of XML documents XML retrieval approaches try to select the most relevant part, represented by an XML element, instead of the whole document. As consequence XML retrieval systems offer two type of query expression, the CO (Content Only) query where user express his needs with simple key word, and the CAS (Content And Structure) query where user can add structural.

Due to the structure of XML document, the traditional RF task becomes more complicated. Indeed, the RF in traditional IR consists in adding the most expressive keywords extracted from of the relevant document In XML retrieval the situation is quite different. The two main questions are:

- First, how to extract from elements, that have different role (semantic), the best terms,

- and the second is how to select the best generative structure that can be added to the query.

In this paper we will present a preliminary work on how one can incorporate the content and the structural information when reformulating the user query. We first give a brief related works in RF and XML retrieval then we present our approach in section 3. The proposed treats the content and the structure separately. In the last section we will describe how we will evaluate our approaches in the framework of INEX.

2. Previous Works

In traditional Information Retrieval, RF consists of reformulating the original query according the user's judgment or automatically said behind RF. It is applied in different model of IR like vector space model presented by Rocchio [7], Tamine [9] has defined the RF in connexionist model, Croft and Haines [1] described RF in an alternative probabilistic model. In XML retrieval the number of RF works is not important. The most works are presented within the framework of the XML retrieval in the company of INEX [2] (Initiative for the Evaluation of XML retrieval).

The working group of V. Mihajlovic and G. Ramirez [6] has proposed a strategy of reformulation applied to the TIJAH [3] model. This last has the same architecture's data bases system. Indeed, the model is composed of three levels: conceptual, logic and physics. At the conceptual level, the authors have adopted the query language Narrowed Extended XPath (NEXI) proposed by INEX in 2003. The logical level is based on the algebra "score area algebra" whose documents were regarded as a continuation of tokens. At the physical level the "MonetDB" ' which was applied to calculate similarity. This last is based to the three measures: tf (frequency of the terms), cf (frequency of the collection) and lp (size with priory). Reformulation is carried out on two stages: the first consists in extract from the document the most relevant elements. This information represents the newspaper where is found the most relevant element, the etiquette of the element and the size which one wishes to find.

Another proposal for a reformulation was presented by the group IBM[4]. This proposal adapted the Rocchio [7]

algorithm to the vector model [5] whose vectors is composed of sub-vectors where each one represents a level of granularity, They applied the method IT (Lexical Affinity) for separation of the relevant documents and not relevant.

3. Relevance Feedback in XML documents

Up until now, in Information Retrieval, the simple keywords are applied in query expansion. But the XML retrieval offers the opportunity to express user's needs by structural information. The main goal of this preliminary work is to present our investigation in CO and CAS queries. More precisely, we discuss how one can introduce structural constraints in the CO query and how one can correct the structural constraints in the CAS query?

These two issues are described separately in the following the subsections.

3.1. Contextual relevance feedback

According to the previous works in traditional Information Retrieval, we have noticed that the more appropriate method to expand query in vector space model is to add weighted key words that represent the most relevant documents and reject those express the irrelevant documents. This method is represented by the formula of Rocchio [7]. In the same way we do not apply any more key words of documents but those of various components of this last.

Our approach is expressed in the following formula:

$$Q' = Q + \sum_{np} C_p - \sum_{nnp} C_{np}$$

With:

- Q : vector of the initial request
- Q' : vector of the new request,
- C_p : (resp. C_{np}): vector of a relevant component (resp. not relevant),
- n_p (resp. n_{np}): component count considered relevant (resp. not relevant).

To apply this method, we have to select the most important key words, so it is clear that if we will add all the representative elements key words, the set of the last will be very enormous and we will have various concepts that can bring noise to the retrieval result. For this reason, we have got more important weight for the key words that are repeated in more one element. The key words weight is proportional to the number of appearances in the

elements judged more relevant.

The CO query represent a simple application of contextual RF, but for CAS query, we have to add the Key words (or reject from) to the most generative structure, we will explain in follow how can reconstitute the most generative structure.

3.2. Structural Relevance Feedback

We have seen that contextual RF is based on additional key words, but in structural RF it is not possible to add structure. Thus we are obliged to look for an appropriate method to reconstitute the generative structure that can help the user to get improvement in retrieval.

Our goal is to define and to reconstitute the appropriate generative structure. We have to notice that the appropriate generative structure should not be the most generative because this last represents the totality of documents. That is why; we have to define the smallest common ancestor (sca).

If we consider the following example, we notice that the XML document is represented as a tree in which, the root is the totality of document. The nodes represent the different elements of various granularities and the leaf nodes are the textual information.

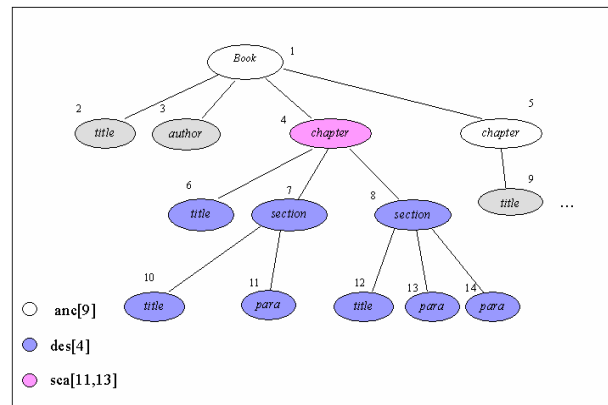


Figure 1: Example of XML document representation

Let us consider the tree structure T,

- Anc[n] is the whole of the ancestors of node n; it is the whole of the nodes which make the way active of the root towards n.
- Des[n] is the whole of the descendants of node n in T; it is the whole of the nodes having n like ancestor.
- Sca (m, n) is the smallest common ancestor of the nodes m and n; it is the first node common to the ways active of m and n towards the root.

If we assume that for given query the IR system returns the nodes 13, 8 and 4, the task is to decide which structure can be introduced in the query: « book/chapter »

or « book/chapter/section » or «book/chapter/section/para ». We notice that « book/chapter » is more generative but the criterion that we have respect in IR is that the information must be exhaustive¹ and specific². So, in our approach, we get more advantage to the structure that is represented by a big number of relevant elements and by considering elements scores. The function that calculates the score 'SScore' of each structure candidate (i.e. which can be injected into the request) is given in follow:

$$SScore = \sum_i^n S_i \cdot \alpha^d$$

With:

S_i score of a relevant element having a joint base with the structure candidate,
 n a number of the relevant judged elements,
 α a constant varying between 0 and 1,
 d is the distance which separates the turned over node, of the last node on the left of the structure candidate.

Example 1

Q is a CO query (Content Only), composed by simple key words: "X, Y",
 We suppose that there are 3 components judged as relevant having respectively the following structures: « /A/B/C », « /A/B/F/L/P » and « /A/B / » and having various weights. It is noticed that structure « /A/B » represents the common factor of the three components, the reformulated request
 Q': will be the query of the type CAS (Content And Structure):
 Q': /A/B [about (X ,Y)].

Example 2

Q is a CAS query expressed in the query language of XFIRM [8]:

//A[about (...X)]//ce: B[Y],

With: A and B are names of the tags of XML documents components. X and Y are key words.
 This query seeks one under component B which contains the key word Y and belongs to the descendants of A in which, one speaks about X. There are 3 components considered to be relevant having respectively, the following structures: « /A/K/C/B », « /A/F/L/B » and « /A/K/B/ » whose corresponding elements have respectively, following weights: 0.5, 0.2 and 0.35.
 We apply thereafter the formula which calculates the score 'SScore' of each structure candidate. The structures

which can be candidates are presented in the following table with their scores ($\alpha=0.8$). We have chosen this value arbitrarily that will be varied in the following experiments.

We have to notice that if α smaller, we give advantage to the more specific structure and if α is bigger, we give advantage to the more generative structure.

/A/K/C/B/	/A/F/L/B/	/A/K/B/	/A/	/A/K/
0.5	0.2	0.35	0.58	0.6

Table 1: Measurement of the candidate structure scores

$$SScore_{/A/K/} = 0.5 \cdot 0.8^2 + 0.35 \cdot 0.8^1 = 0.6$$

$$SScore_{/A/} = 0.5 \cdot 0.8^3 + 0.2 \cdot 0.8^3 + 0.35 \cdot 0.8^2 = 0.58$$

According to this table, we notice that the structure which can be inserted is: « /A/K/ ».

To introduce it into the structured request we use the function of aggregation already used for the made up CAS query according to model of XFIRM [8].

We have to notice that if the structure having the most important weight is the same of the initial structure query we consider in aggregation the structure on the second rank.

If we suppose that N and M are two different elements: The node result of aggregation (N and M) and its relevance are represented by the pair (l, r_1) . L is the ancestor nearest is:

$$r_1 = \text{aggr}_{\text{and}}(r_n, r_m, \text{dist}(l,n), \text{dist}(l,m))$$

With:

$$\text{aggr}_{\text{and}}(r_n, r_m, \text{dist}(l,n), \text{dist}(l,m)) = \frac{r_n / \text{dist}(l,n) + r_m / \text{dist}(l,m)}$$

$\text{dist}(x, y)$ is the distance which separates X and y in-depth and r_i the value of relevance of element i.

The final reformulated query is the result of aggregated structure where content condition is the initial keyword added with expansion given by Contextual RF.

4. Experiments

Application of reformulation is applied on XFIRM. It is a Flexible Model of Information Retrieval for the

¹ An element is judged exhaustive if it involves the all information needed by the user.

² An element is judged specific if the all information included is related to the subject of the user's query.

storage and the interrogation of documents XML prepared within our team. It is based on data storage and a simple query language, allowing the user to formulate his need using simple key words or in a more precise way by integrating constraints structure of the documents. The similarity measure is based on *tf* (term frequency) and *ief* (inverse element frequency).

To evaluate the results of our contribution, we have resorted to the company of INEX (Initiative for the Evaluation of XML Retrieval) [2]. The purpose of this company is to be able to evaluate the XML Retrieval systems by providing test collections of XML documents, the procedures of evaluation and a forum. This company allows to the participating organizations to compare their results. Collections of the test evaluation XML retrieval treat the elements of various granularities. The corpus is composed of papers coming from IEEE Computer Society marked out with format XML; they constitute a collection from approximately 750 MB, containing more than 13000 articles published between 1995 and 2004, coming from 21 reviews. An average article is composed of approximately 1500 nodes XML. The evaluation is based on the two criteria: exhaustiveness and specificity. It is the participant's verdict which will decide the degree of tow criteria.

We have carried out our approach that will be evaluated on INEX 2005. The ultimate result will be given in November 2005 and since it is our first participation in RF task, we have not yet had an official result.

5. Conclusion

In this paper, we have presented our search work done in the XML Retrieval. Our work represents a new approach in the Relevance Feedback task which we have applied a new strategy of contextual expansion query and our proposition is to reconstitute the appropriate generative structure in order to get the most exhaustive and specific information. In future, we will evaluate our approaches in INEX 2005.

6. References

- [1] W. Croft and D. Harper. Using probabilistic models of information retrieval without relevance information. *Journal of Documentation*. 35(4): 285_295, 1979.
- [2] INEX 2004 Workshop Pre-Proceedings. <http://inex.is.informatik.uni-duisburg.de:2004/>
- [3] J. A. List, V. Mihajlovic, A. P. de Vries and G. Ramirez. The TIJAH XML-IR system at INEX 2003 (DRAFT. *Proceedings of INEX 2003 Workshop*:

102_109, 2003.

- [4] Y. Mass and M. Mandelbrod. Relevance Feedback for XML Retrieval. *INEX 2004 Workshop Pre-Proceedings*: 154_157, 2004.

- [5] M. Mass, M. Mandelbrod, E. Amitay, Y. Maarek. and A. Soffer. JuruXML-an XML retrieval system at INEX'02. *Proceedings of the First Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)*: 73_80, 2002.

- [6] V. Mihajlovic, G. Ramirez, A. de Vries. and D. Hiemstra. TIJAH at INEX 2004 Modeling Phrases and Relevance Feedback. *INEX 2004 Workshop Pre-Proceedings*: 141_148, 2004.

- [7] J. J. Rocchio. Relevance feedback in information retrieval. *The SMART retrieval system - experiments in automatic document processing*: 313_323, 1971.

- [8] K. Sauvagnat, M. Boughanem and C. Chrisment. Searching XML documents using relevance propagation. *SPIRE 04*: 242_254, 2004.

- [9] L. Tamine and M. Boughanem: Query Optimization Using An Improved Genetic Algorithm. *CIKM 2000*: 368-373, 2000.

- [10] Extensible markup language (XML). <http://www.w3.org/TR/1998/REC-xml-19980210>.